match a query subject or category. The process extends the Training Set to include all documents pointing to or pointed to by each document in the Training Set. Using information that describes the links between the documents, the process seeks the best Authorities and Hubs that match the query or category. Mathematically, the Authorities and Hubs are the principal Eigenvectors of matrices representing the link relations between the documents.

In another approach, the GOOGLE project (at URL google.stanford.edu) uses a process of generating PageRanks. PageRanks are iteratively updated based on linked hypertext structures. The resulting PageRanks measure the general connectedness of documents on the Web, without regard to a particular category or query. The assumption is that more connected documents will tend to be of general interest to most users.

Replace the paragraph on page 12, line 17 with:

For instance, if one document is clicked often in the same time period as another, and both documents are viewed for a long period of time, the classification system may determine that the two pages are relevant to the user's interests and have Similarity. Two assumptions are made. First, the frequency at which a user changes his or her topic of interest is much slower than the frequency at which a user changes pages. Second, a user's interest in a page can be estimated by a function of the time that the user spends in viewing the page.

Replace the paragraph on page 14, line 3 with:

5. TITLE SIMILARITY

Replace the paragraph on page 14, line 16 with:

6. URL SIMILARITY

Replace the paragraph on page 15, line 4 with:

7. CACHE HIT LOG SIMILARITY

Replace the paragraph beginning at page 15, line 20 with:

A human user of the client 200, or an agent executing in the client, instructs browser 202 to request a hypertext document according to a particular location identifier. For

example, a Web browser of the client may request a Web document using its URL, such as "www.inktomi.com/". Browser 202 submits the request to cache server 208. The cache server 208 determines whether the requested document is already in the cache 210. If it is, the cache server 208 delivers the requested document to the browser 202 from the cache 210. If it is not, cache server 208 uses a domain name service or similar network element of network 204 to determine the location of origin server 220. Cache server 208 then requests the document from origin server 220, via network 204. Finally, cache server 208 stores a copy of the document in cache 210, and passes a copy of the document back to browser 202. Thus, normally, all Web traffic directed from browser 202 passes through the cache server 208. The cache server is thereby in the ideal position to log users' requests for various Web documents.

Replace the paragraphs beginning at page 16, line 23 with:

14-Feb-1999 08:01:22, 255.1.2.254, "www.inktomi.com/index.html"

14-Feb-1999 08:01:23, 199.22.131.44, "www.mwe.com/"

14-Feb-1999 08:01:26, 255.1.2.254, "www.inktomi.com/products"

14-Feb-1999 08:01:27, 199.22.131.44, "www.mwe.com/bios"

14-Feb-1999 08:01:31, 255.1.2.254, "www.inktomi.com/products/trafficserver.html"

14-Feb-1999 08:01:32, 199.22.131.44, "www.mwe/com/bios/palec.htm"

Replace the abstract at page 40, line 2 with the following:

A method and apparatus are provided for determining when electronic documents stored in a large collection of documents are similar to one another. A plurality of similarity information is derived from the documents. The similarity information may be based on a variety of factors, including hyperlinks in the documents, text similarity, user click-through information, similarity in the titles of the documents or their location identifiers, and patterns of user viewing. The similarity information is fed to a combination function that synthesizes the various measures of similarity information into combined similarity information. Using